# StreptomeDB 2.0 - An extended resource of natural products produced by streptomycetes

Klementz, Dennis; Döring, Kersten; Lucas, Xavier; Telukunta, Kiran K.; Erxleben, Anika; Deubel, Denise

*Published in:*
Nucleic Acids Research

*DOI:*
10.1093/nar/gkv1319

*Publication date:*
2016

*Published online 28 November 2015* *Nucleic Acids Research, 2016, Vol. 44, Database issue* ***D509–D514***
*doi: 10.1093/nar/gkv1319*

# StreptomeDB 2.0—an extended resource of natural products produced by streptomycetes

**Dennis Klementz[1,†], Kersten Döring[1,†], Xavier Lucas[1,2], Kiran K. Telukunta[1], Anika Erxleben[1,3], Denise Deubel[4], Astrid Erber[4], Irene Santillana[4], Oliver S. Thomas[1], Andreas Bechthold[4] and Stefan Günther[1,*]**

[1]Pharmaceutical Bioinformatics, Institute of Pharmaceutical Sciences, Albert-Ludwigs-University, Hermann-Herder-Strasse 9, Freiburg 79104, Germany, [2]School of Life Sciences, Division of Biological Chemistry and Drug Discovery, University of Dundee, James Black Centre, Dow Street, Dundee DD1 5EH, UK, [3]Chair for Bioinformatics, Department of Computer Science, University of Freiburg, Georges-Koehler-Allee 106, Freiburg 79110, Germany and [4]Pharmaceutical Biology, Institute of Pharmaceutical Sciences, Albert-Ludwigs-University, 79104 Freiburg, Germany

Received September 15, 2015; Revised October 28, 2015; Accepted November 10, 2015

## ABSTRACT

**Over the last decades, the genus *Streptomyces* has stirred huge interest in the scientific community as a source of bioactive compounds. The majority of all known antibiotics is isolated from these bacterial strains, as well as a variety of other drugs such as antitumor agents, immunosuppressants and antifungals. To the best of our knowledge, StreptomeDB was the first database focusing on compounds produced by streptomycetes. The new version presented herein represents a major step forward: its content has been increased to over 4000 compounds and more than 2500 host organisms. In addition, we have extended the background information and included hundreds of new manually curated references to literature. The latest update features a unique scaffold-based navigation system, which enables the exploration of the chemical diversity of StreptomeDB on a structural basis. We have included a phylogenetic tree, based on 16S rRNA sequences, which comprises more than two-thirds of the included host organisms. It enables visualizing the frequency, appearance, and persistence of compounds and scaffolds in an evolutionary context. Additionally, we have included predicted MS- and NMR-spectra of thousands of compounds for assignment of experimental data. The database is freely accessible via http://www.pharmaceutical-bioinformatics.org/streptomedb.**

## INTRODUCTION

The growing number of annotated natural products (NPs) in databases, such as Super Natural II (∼326 000 compounds; 2D structures; physicochemical properties; predicted toxicity class; potential vendors) (1), KNApSAcK (∼51 000 metabolites; ∼111 000 metabolite-species pairs) (2), UNPD (∼229 000 compounds; 3D structures) (3) or NORINE (∼1200 NRPs) (4) reflects the increasing interest in these molecules. StreptomeDB is, to the best of our knowledge, the biggest compilation of NPs produced by streptomycetes. Its content has been collected from thousands of abstracts and full papers by text mining methods and extensive manual curation. In addition, it includes comprehensive background information such as host organisms, predicted physicochemical properties, synthesis routes and biological activities.

*Streptomyces* is a genus of Gram-positive actinobacteria. They can be found all around the world in soil samples and successfully inhabit many terrestrial and aquatic niches (5). Since the discovery of streptomycin by Albert Schatz in the group of Selman Waksman in 1943 (6), they have become one of the best studied bacterial *genera*. Over 60% of all known antibiotics have been isolated from streptomycetes (5). Many of these compounds are approved drugs, such as the well-known agents tetracycline, daptomycin and chloramphenicol (7). Besides antibiotics, there is a rich diversity of other secondary metabolites with a plethora of different biological activities and therapeutic potential exclusively produced by streptomycetes. Blockbuster drugs like the anti-parasitic agent avermectin, the immunosuppressant rapamycin or the lipase inhibitor lipstatin are just a few examples (8,9,10). There is also a large number of potential antibiotics or other bioactive compounds found in Strep-

*To whom correspondence should be addressed. Tel: +49 761 203 4871; Fax: +49 761 203 97769; Email: stefan.guenther@pharmazie.uni-freiburg.de
†These authors contributed equally to the paper as first authors.

tomeDB which are not in clinical use (11). As the recent development of highly effective griselimycin analogs shows (12), these molecules can be considered as possible cornerstones for the future development of semi-synthetic drugs (13).

A common starting point of a modern drug discovery campaign is an *in silico* high-throughput screening for small molecules or fragments that can interact with a therapeutic target of interest (14). The method strongly relies on high quality and diverse compound libraries like StreptomeDB. Although NPs are in many cases harder to obtain than compounds produced by chemical synthesis and are rarely found in vendors' catalogs, they are nonetheless a valuable addition to screening libraries (11,13). Due to their complex stereochemistry, they are often more selective than synthetic compounds and are particularly suitable–for addressing low-druggable targets (15,16). Furthermore, NPs are in most cases produced enantiomerically pure, whereas enantiomeric purification of synthetic compounds can be an expensive and exhausting task (17).

Due to the variety of chemical precursors and diversity of produced scaffolds, streptomycetes are versatile hosts for industrial heterologous expression (18). Developments in biotechnology and synthetic biology have led to artificially minimized host strains that can be utilized for the production of various compounds (19). For example, the genetically engineered SUKA strains, descendants of the industrial microorganism *Streptomyces avermitilis*, which is already known for its highly efficient production of the anthelmintic agent avermectin (20), have proven useful as a heterologous host in the production of several secondary metabolites. Beyond the production of bioactive compounds, the rich diversity of chemical scaffolds produced by streptomycetes enables their use as precursors for many semi-synthetic approaches. For example, *Streptomyces venezuelae* has recently been employed as a producer of bisabolenes (21), which are sesquiterpenes present in several essential oils in plants. The advanced biofuel bisabolane (22) is synthesized from them in a single step by chemical hydrogenation (23).

Although StreptomeDB demonstrates that there are already thousands of compounds isolated from streptomycetes, their potential as a source of NPs is far from exhausted (24). Advancing methods still allow for the discovery of new strains and compounds, such as the phenalinolactones, i.e. terpene glycosides with a rare, highly oxidized γ-butyrolactone structure (25).

The broad interest in streptomycetes has so far led to more than 23 000 publications in PubMed. In the last years, the number of publications per year was constantly growing to reach about 1000 in 2014, stressing the need for an updated StreptomeDB that makes this vast amount of data accessible.

Here we present StreptomeDB 2.0, a major update with an increase from about 2500 to over 4000 compounds, hundreds of which required manual drawing prior to insertion into the database, and more than 2500 host strains. Additionally, the introduced new features assist in the navigation through the vast chemical diversity produced by streptomycetes (Figure 1). This allows for a deeper understanding of the complexity and evolution of the synthesis machinery that makes this *genus* so attractive for academic and industrial research.

## MATERIALS AND METHODS

### Data collection

Manual curation of literature was carried out following the protocol of the original database (11). For the sake of completeness, special focus was put on gathering compounds from full texts. Additionally, hundreds of compounds with name not provided in the referenced manuscript or not found in PubChem were manually depicted prior to insertion into the database.

### Generation of the phylogenetic tree

The phylogenetic tree is based on the most common 16S rRNA sequence per strain. Over 20 000 sequences from freely available sources, such as silva (26) and ENA (27), were filtered for integrity and length and mapped to organisms in StreptomeDB. This resulted, in some cases, in over 150 different 16S rRNA sequences for a single strain. If it was not possible to identify the best sequence, a hypothetical, most common sequence was derived. For this approach, the strain-specific datasets were first scanned with BLAST+ (28) for the sequences that produce the most hits with an *E*-value below $1 \times 10^{-4}$ and a coverage over 97%. If this resulted in more than one sequence, a consensus sequence was built with the dumb_consensus function, as implemented in Biopython (29). Sequences were aligned with ClustalW (30). Phylogenetic analysis of the resulting alignment was performed with the MEGA software package (31) using the maximum likelihood and maximum parsimony algorithms, each with 250 bootstrap replications. Sub-clusters in the tree were detected by summarizing all leafs that have a common ancestor node within range of 0.07 nt substitutions per side, using the DendroPy library (32). The final editing and visualization was done with the ETE 2 toolkit (33).

### Gene clusters, genomes and taxonomy

If available, links to gene clusters in DoBISCUIT (34) or MiBIG(35) are provided in StreptomeDB. Host organisms are linked to freely available full genomes from GenBank (36) and entries in the NCBI taxonomy database (37). Substrains and mutants without taxonomy ID are linked to their primary strains.

### Scaffold-based molecular decomposition

The Scaffold Decomposition tool included in Canvas 2.3 (Schrödinger, LLC, New York, NY, USA) was used to process molecules from StreptomeDB and extract all represented scaffolds. Small molecules can comprise one or more level 0 scaffolds, i.e. root cyclic or polycyclic independent structures. Two chemically equivalent or different level 0 entities connected by an aliphatic or functionalized linker define a level 1 scaffold. Subunits composed of a level 0 plus a level 1 scaffold generate a level 2 scaffold, and so on (38).

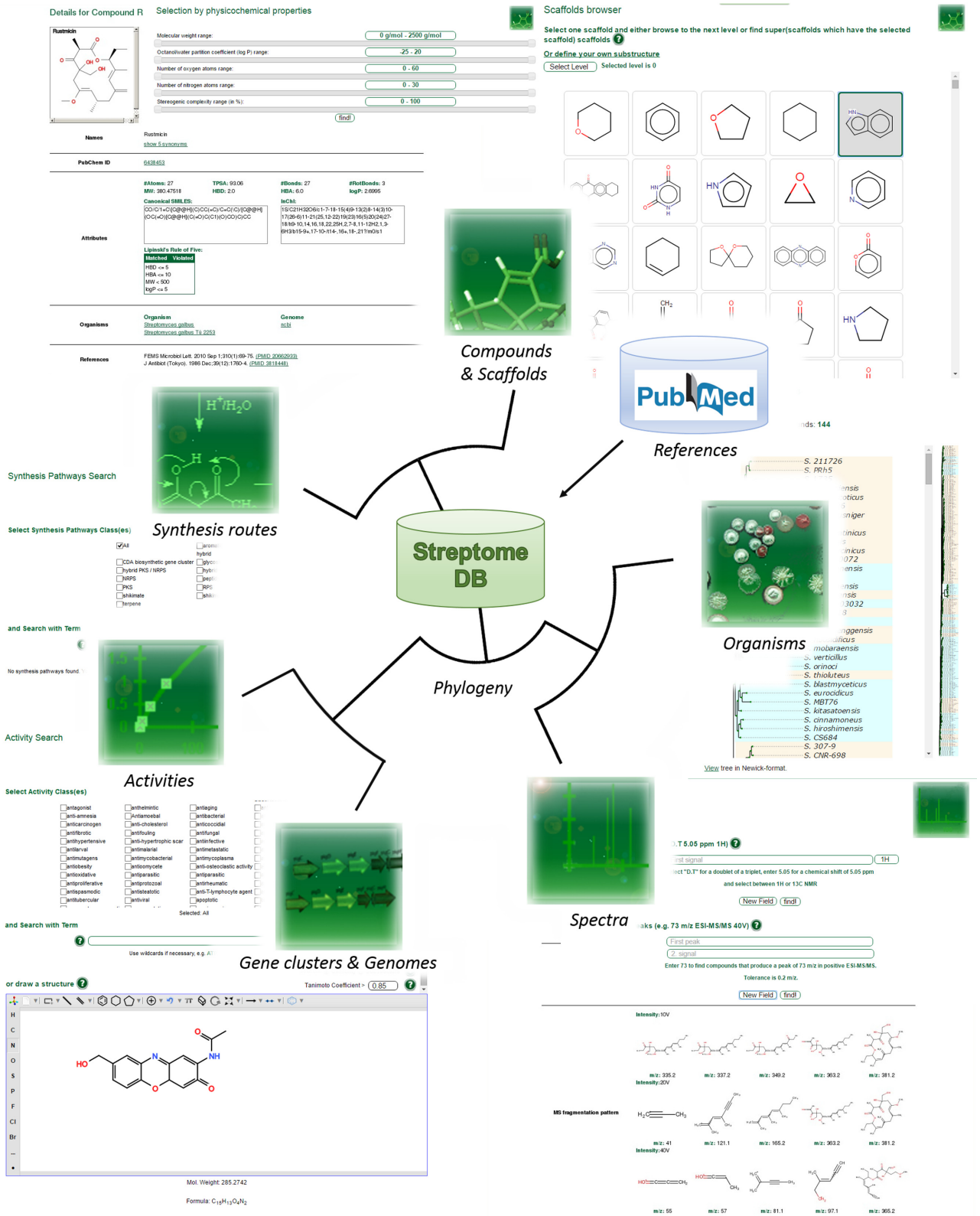


**Figure 1.** Features and search options implemented in StreptomeDB.

### Prediction of fragmentation patterns in mass spectra

The software CFM-ID, based on competitive fragmentation modeling to produce a probabilistic generative model for electrospray tandem spectrometry (ESI-MS/MS) fragmentation (39), was used to predict the fragmentation spectrum of molecules with ≤30 heavy atoms in positive ionization mode at 10, 20 and 40 V, and assign the resulting peaks to their chemical structure. For each intensity level, the most intense peaks (max. 5) are reported in StreptomeDB.

### Prediction of $^{1}H$ and $^{13}C$ NMR peaks

The nuclear magnetic resonance (NMR) Predictor tool, as implemented in the command-line platform cxcalc (Marvin 15.4.13.0, 2015, ChemAxon, http://www.chemaxon.com), was used to predict the $^{1}H$ and $^{13}C$ NMR spectra of all molecules.

## NEW FEATURES AND UPDATES

### Phylogenetic tree

A new feature of StreptomeDB is the comprehensive phylogenetic tree based on 16S rRNA sequences. Considering the usually similar six or more 16S rRNA (40) sequences present in a *Streptomyces* genome (and in some cases several similar sequencing attempts for one sequence, which cannot be distinguished in quality or length), we generated a consensus sequence for each strain. We have chosen a tree based on 16S rRNA to achieve the highest possible coverage of the data contained in StreptomeDB. For a better overview and because of the relatively short evolutionary distance, more than 1200 sub strains are represented by 340 parent stains, e.g. clicking on *'Streptomyces griseus'* leads to 128 compounds produced either by *S. griseus* itself or one of its 58 sub strains, such as *S. griseus* or *S. griseus var. psychrophilus*. The apparently small number of 340 shown strains comprises more than two-thirds of the compounds in the database, therefore indicating a good coverage of the available molecules. Additionally, it is possible to access the phylogenetic tree from any search, with the producing organisms of the related compounds highlighted. This provides the possibility to visualize the distribution of a certain scaffold, chemotype, compound, bioactivity or synthetic route in an evolutionary context.

### Scaffold-based navigation and search system

A new scaffold-based navigation has been integrated and the compound search system has been substantially improved. All molecules have been redefined by means of their scaffold framework (38), and there is now the possibility to browse through the gathered chemical scaffolds. At each scaffold level (*Search -> Scaffolds browser*), the user can choose one or several scaffolds, which are sorted by frequency among the gathered compounds, and display either compounds or higher-level scaffolds. A phylogenetic tree with highlighted producing organisms can now be easily accessed. Furthermore, compound cards contain a list of represented scaffolds, which in turn link to a list of compounds that contain the scaffold. This enables the bidirectional browsing through the chemical diversity of the database on a structural basis and its connection to the evolution of the producing organisms. Table 1 shows that StreptomeDB contains more than 1000 level 0 scaffolds and hundreds of higher-level scaffolds, including 281 scaffolds of very advanced framework and high molecular weight. In addition, searches based on user-defined chemical structure patterns can be performed taking advantage of fast substructure searches (Search -> Compound(structure) section).

### Additional content

Compound cards now provide predicted MS and NMR spectra for thousands of compounds. The MS/MS spectra include the five most intensive peaks for 10, 20 and 40 V energy levels. The chemical structure of each peak is depicted. NMR data are listed in a table that includes intensity, multiplicity and chemical shift for both, $^{1}H$ and $^{13}C$ NMR. It is possible to query the database for any combination of $^{1}H$ and $^{13}C$ signals or MS peaks. Additionally, there is the possibility to query the database by physicochemical properties, including molecular weight, octanol/water partition coefficient (*c*log $P_{o/w}$), number of nitrogen and oxygen atoms and stereogenic complexity, i.e. $C^*/C_T$ (16). Finally, StreptomeDB includes links to all publicly available gene clusters and genomes for compounds and host organisms.

### Back end and library updates

All software libraries related to the web page, its database back end and the in-house curation platform were updated to the latest versions. The structures of curated molecules were mapped to canonical SMILES using ChemicalToolBoX (41). The update steps have been redesigned to upload new molecules and related data to the database back end automatically, based on the uniqueness of structures.

## CONCLUSION AND FUTURE PROSPECTS

Here we present a major update on StreptomeDB, which implements a series of new features to offer easier access to the huge amount of data related to streptomycetes, which is hidden in literature. Furthermore, StreptomeDB improves the integration of this information in a chemical or evolutionary context.

From a chemical point of view, the implementation of a comprehensive scaffold-based navigation system enables browsing the rich diversity of NPs produced by streptomycetes. Scaffold-based molecular representations are a prominent tool in medicinal chemistry (38) and allow for classifying and comparing the coverage and content of chemical libraries. We could identify hundreds of naturally occurring level 1 scaffolds not found in purchasable compounds (15). This is of particular interest in drug discovery, as on average each commercialized drug contains a novel scaffold, thus stressing the crucial relevance of mining both natural sources and literature for molecules. Clearly, secondary metabolites produced by streptomycetes are an attractive source of chemical diversity.

The phylogenetic tree enables the visualization of the distribution of these scaffolds and the compounds they represent. Especially in projects that deal with poorly studied

**Table 1.** Chemical diversity of StreptomeDB

| Scaffold level | No. of unique scaffolds | Scaffold level | No. of unique scaffolds |
|---|---|---|---|
| 0 | 1,032 | 6 | 213 |
| 1 | 732 | 7 | 137 |
| 2 | 672 | 8 | 96 |
| 3 | 559 | 9 | 91 |
| 4 | 469 | 10 | 84 |
| 5 | 314 | >10 | 281 |

Scaffold levels and the number of unique scaffolds included in the database.

strains, this feature offers an easy access to additional references and background information of closely related organisms. It is a new approach toward the understanding of the evolution of secondary metabolism. Together with the offered gene clusters and genomes it provides all building blocks for comparative genome analysis, reconstruction of metabolic networks, or building a customized, more specialized phylogenetic tree. Motivated by the decreasing costs and massive efforts that are currently undertaken in genome sequencing, the number of published genomes and gene clusters is quickly increasing (42). Therefore, we are looking forward to expanding these features to an even greater extent.

We have also improved the characterization and querying of compounds by including predicted MS and NMR data. These techniques are widely used in structural determination of secondary metabolites, and we realized that a tool for comparing experimental MS spectra with that of compounds in the database would be extremely beneficial for the users of StreptomeDB. Now this is possible by querying the database with experimental peaks, either from MS or NMR determination, to retrieve compounds with matching patterns.

Due to the increasing popularity of StreptomeDB, we want to encourage users to contact us if they find missing data or have optimization proposals.

In conclusion, StreptomeDB is a versatile platform for the gathering of information for projects that deal with streptomycetes. It stands out from other popular NP databases with a broader focus, such as Supernatural II, KNAPsACK or UNPD (1,2,3): built on top of a highly networked structure, StreptomeDB facilitates exposing coherences between different data, such as scaffolds, activities or phylogenetic distribution. It offers an ideal starting point and a vast amount of supplementary resources for the discovery of new compounds, virtual screening campaigns, biochemical engineering and cheminformatics.

## AVAILABILITY

StreptomeDB is freely available, open to all users, and has no login requirements. All compounds can be downloaded with metadata in SD-Format at http://www.pharmaceutical-bioinformatics.org/streptomedb/download.

## ACKNOWLEDGEMENTS

We thank Michael Becer, Sven Enderle and Anna Hähnlein for assistance in literature curation. Additionally, we want to thank all users who contributed to the development of StreptomeDB by giving us helpful suggestions and critical feedback.

## FUNDING

German National Research Foundation (DFG, LIS45, 1225/3-1 and RTG 1976). Funding for open access charge: DFG Research Training Group 1976.
*Conflict of interest statement.* None declared.

## REFERENCES

1. Banerjee,P., Erehman,J., Gohlke,B.O., Wilhelm,T., Preissner,R. and Dunkel,M. (2015) Super Natural II–a database of natural products. *Nucleic Acids Res.*, **43**, D935–D939.
2. Nakamura,Y., Afendi,F.M., Parvin,A.K., Ono,N., Tanaka,K., Morita,A., Sato,T., Sugiura,T., Altaf-Ul-Amin,M. and Kanaya,S. (2013) KNApSAcK metabolite activity database for retrieving the relationships between metabolites and biological activities. *Plant Cell Physiol.*, **55**, e7.
3. Gu,J., Gui,Y., Chen,L., Yuan,G., Lu,H.Z. and Xu,X. (2013) Use of natural products as chemical library for drug discovery and network pharmacology. *PLoS One*, **8**, e62839.
4. Caboche,S., Pupin,M., Leclère,V., Fontaine,A., Jacques,P. and Kucherov,G. (2008) NORINE: a database of nonribosomal peptides. *Nucleic Acids Res.*, **36**, D326–D331.
5. Hopwood,D.A. (2006) Soil to genomics: the *Streptomyces* chromosome. *Annu. Rev. Genet.*, **40**, 1–23.
6. Zetterström,R. (2007) Selman A. Waksman (1888–1973) Nobel Prize in 1952 for the discovery of streptomycin, the first antibiotic effective against tuberculosis. *Acta Paediatr.*, **96**, 317–319.
7. Procópio,R.E., Silva,I.R., Martins,M.K., Azevedo,J.L. and Araújo,J.M. (2012) Antibiotics produced by *Streptomyces*. *Braz. J. Infect. Dis.*, **16**, 466–471.
8. Kose,L.P., Gülçin,İ., Özdemir,H., Atasever,A., Alwasel,S.H. and Supuran,C.T. (2015) The effects of some avermectins on bovine carbonic anhydrase enzyme. *J. Enzyme Inhib. Med. Chem.*, doi:10.3109/14756366.2015.1064406.
9. Krenzien,F., ElKhal,A., Quante,M., Rodriguez Cetina Biefer,H., Hirofumi,U., Gabardi,S. and Tullius,S.G. (2015) A rationale for age-adapted immunosuppression in organ transplantation. *Transplantation*, **99**, 2258–2268.
10. Bai,T., Zhang,D., Lin,S., Long,Q., Wang,Y., Ou,H., Kang,Q., Deng,Z., Liu,W. and Tao,M. (2014) Operon for biosynthesis of lipstatin, the beta-lactone inhibitor of human pancreatic lipase. *Appl. Environ. Microbiol.*, **80**, 7473–7483.
11. Lucas,X., Senger,C., Erxleben,A., Grüning,B.A., Döring,K., Mosch,J., Flemming,S. and Günther,S. (2013) StreptomeDB: a resource for natural compounds isolated from *Streptomyces* species. *Nucleic Acids Res.*, **41**, D1130–D1136.
12. Kling,A., Lukat,P., Almeida,D.V., Bauer,A., Fontaine,E., Sordello,S., Zaburannyi,N., Herrmann,J., Wenzel,S.C., König,C. *et al.* (2015) Targeting DnaN for tuberculosis therapy using novel griselimycins. *Science*, **348**, 1106–1112.
13. Butler,M.S. and Buss,A.D. (2006) Natural products–the future scaffolds for novel antibiotics? *Biochem. Pharmacol.*, **71**, 919–929.

14. Karthikeyan,M. and Vyas,R. (2015) Role of open source tools and resources in virtual screening for drug discovery. *Comb. Chem. High Throughput Screen.*, **18**, 528–543.
15. Lucas,X., Grüning,B.A., Bleher,S. and Günther,S. (2015) The purchasable chemical space: a detailed picture. *J. Chem. Inf. Model.*, **55**, 915–924.
16. Lucas,X. and Günther,S. (2014) Using chiral molecules as an approach to address low-druggability recognition sites. *J. Comput. Chem.*, **35**, 2114–2121.
17. Ribeiro,A.R., Maia,A.S., Cass,Q.B. and Tiritan,M.E. (2014) Enantioseparation of chiral pharmaceuticals in biomedical and environmental analyses by liquid chromatography: an overview. *J. Chromatogr. B. Analyt. Technol. Biomed. Life Sci.*, **968**, 8–21.
18. Yamada,Y., Arima,S., Nagamitsu,T., Johmoto,K., Uekusa,H., Eguchi,T., Shin-ya,K., Cane,D.E. and Ikeda,H. (2015) Novel terpenes generated by heterologous expression of bacterial terpene synthase genes in an engineered *Streptomyces* host. *J. Antibiot. (Tokyo).*, **68**, 385–394.
19. Komatsu,M., Komatsu,K., Koiwai,H., Yamada,Y., Kozone,I., Izumikawa,M., Hashimoto,J., Takagi,M., Omura,S., Shin-ya,K. *et al.* (2013) Engineered *Streptomyces avermitilis* host for heterologous expression of biosynthetic gene cluster for secondary metabolites. *ACS Synth. Biol.*, **2**, 384–396.
20. Liu,W., Zhang,Q., Guo,J., Chen,Z., Li,J. and Wen,Y. (2015) Increasing avermectin production in *Streptomyces avermitilis* by manipulating the expression of a novel TetR-family regulator and its target gene product. *Appl. Environ. Microbiol.*, **81**, 5157–5173.
21. Phelan,R.M., Sekurova,O.N., Keasling,J.D. and Zotchev,S.B. (2015) Engineering terpene biosynthesis in *Streptomyces* for production of the advanced biofuel precursor bisabolene. *ACS Synth. Biol.*, **4**, 393–399.
22. Beller,H.R., Lee,T.S. and Katz,L. (2015) Natural products as biofuels and bio-based chemicals: fatty acids and isoprenoids. *Nat. Prod. Rep.*, **32**, 1508–1526.
23. Peralta-Yahya,P.P., Ouellet,M., Chan,R., Mukhopadhyay,A., Keasling,J.D. and Lee,T.S. (2011) Identification and microbial production of a terpene-based advanced biofuel. *Nat. Commun.*, **2**, doi:10.1038/ncomms1494.
24. Dhakal,D. and Sohng,J.K. (2015) Commentary: Toward a new focus in antibiotic and drug discovery from the *Streptomyces* arsenal. *Front. Microbiol.*, **6**, doi:10.3389/fmicb.2015.00727.
25. Kiske,C., Erxleben,A., Lucas,X., Willmann,L., Klementz,D., Günther,S., Römer,W. and Kammerer,B. (2014) Metabolic pathway monitoring of phenalinolactone biosynthesis from *Streptomyces* sp. Tü6071 by liquid chromatography/mass spectrometry coupling. *Rapid Commun. Mass Spectrom.*, **28**, 1459–1467.
26. Quast,C., Pruesse,E., Yilmaz,P., Gerken,J., Schweer,T., Yarza,P., Peplies,J. and Glöckner,F.O. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.*, **41**, D590–D596.
27. Cochrane,G., Alako,B., Amid,C., Bower,L., Cerdeño-Tárraga,A., Cleland,I., Gibson,R., Goodgame,N., Jang,M., Kay,S. *et al.* (2013) Facing growth in the European Nucleotide Archive. *Nucleic Acids Res.*, **41**, D30–D35.
28. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, doi:10.1186/1471-2105-10-421.
29. Cock,P.J., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F., Wilczynski,B. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
30. McWilliam,H., Li,W., Uludag,M., Squizzato,S., Park,Y.M., Buso,N., Cowley,A.P. and Lopez,R. (2013) Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Res.*, **41**, W597–W600.
31. Tamura,K., Stecher,G., Peterson,D., Filipski,A. and Kumar,S. (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.*, **30**, 2725–2729.
32. Sukumaran,J. and Holder,M.T. (2010) DendroPy: a Python library for phylogenetic computing. *Bioinformatics*, **26**, 1569–1571.
33. Huerta-Cepas,J., Dopazo,J. and Gabaldón,T. (2010) ETE: a python Environment for Tree Exploration. *BMC Bioinformatics*, **11**, doi:10.1186/1471-2105-11-24.
34. Ichikawa,N., Sasagawa,M., Yamamoto,M., Komaki,H., Yoshida,Y., Yamazaki,S. and Fujita,N. (2013) DoBISCUIT: a database of secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res.*, **41**, D408–D414.
35. Medema,M.H., Kottmann,R., Yilmaz,P., Cummings,M., Biggins,J.B., Blin,K., de Bruijn,I., Chooi,Y.H., Claesen,J., Coates,R.C. *et al.* (2015) Minimum information about a biosynthetic gene cluster. *Nat. Chem. Biol.*, **11**, 625–631.
36. Benson,D.A., Cavanaugh,M., Clark,K., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2013) GenBank. *Nucleic Acids Res.*, **41**, D36–D42.
37. Sayers,E.W., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **37**, D5–D15.
38. Langdon,S.R., Brown,N. and Blagg,J. (2011) Scaffold diversity of exemplified medicinal chemistry space. *J. Chem. Inf. Model.*, **51**, 2174–2185.
39. Allen,F., Greiner,R. and Wishart,D. (2014) Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics*, **11**, 98–110.
40. Coenye,T. and Vandamme,P. (2003) Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. *FEMS Microbiol. Lett.*, **228**, 45–49.
41. Lucas,X., Grüning,B.A. and Günther,S. (2014) ChemicalToolBoX and its application on the study of the drug like and purchasable space. *J. Cheminform.*, **6**, P51.
42. Land,M., Hauser,L., Jun,S.R., Nookaew,I., Leuze,M.R., Ahn,T.H., Karpinets,T., Lund,O., Kora,G., Wassenaar,T. *et al.* (2015) Insights from 20 years of bacterial genome sequencing. *Funct. Integr. Genomics*, **15**, 141–161.